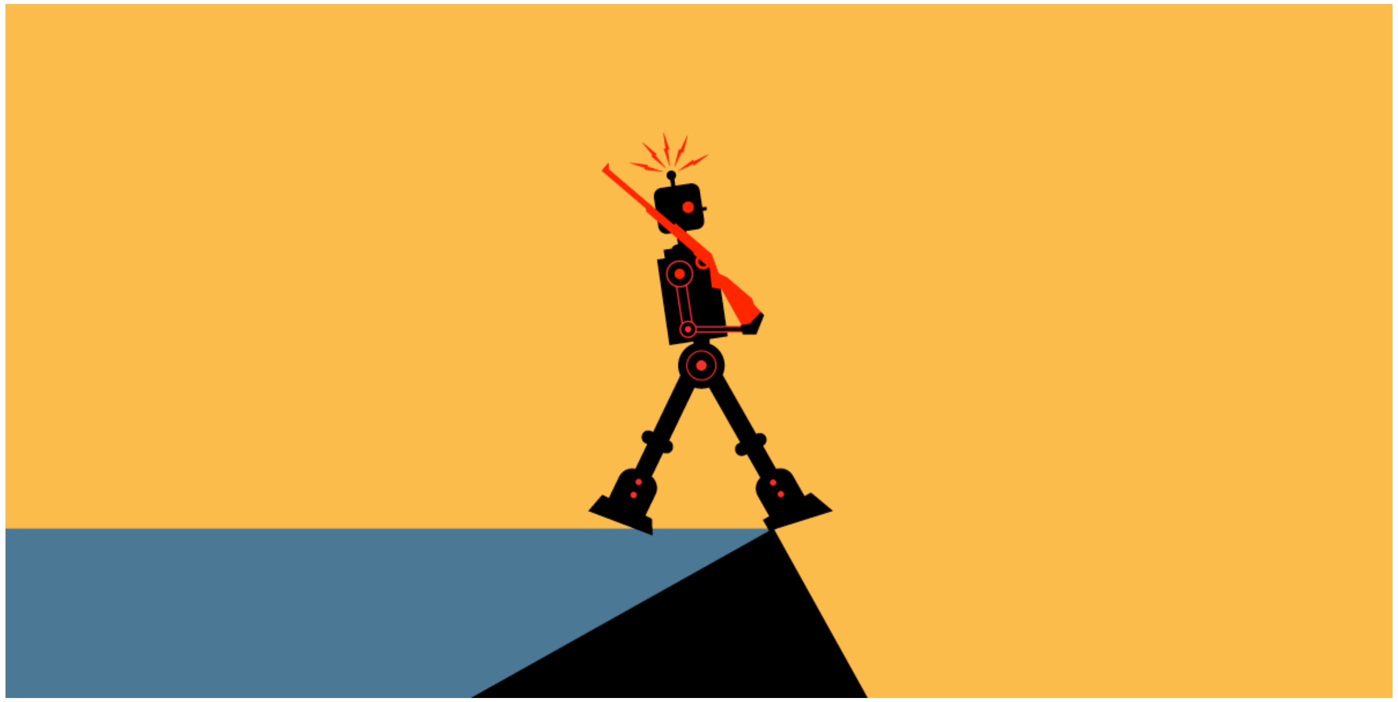
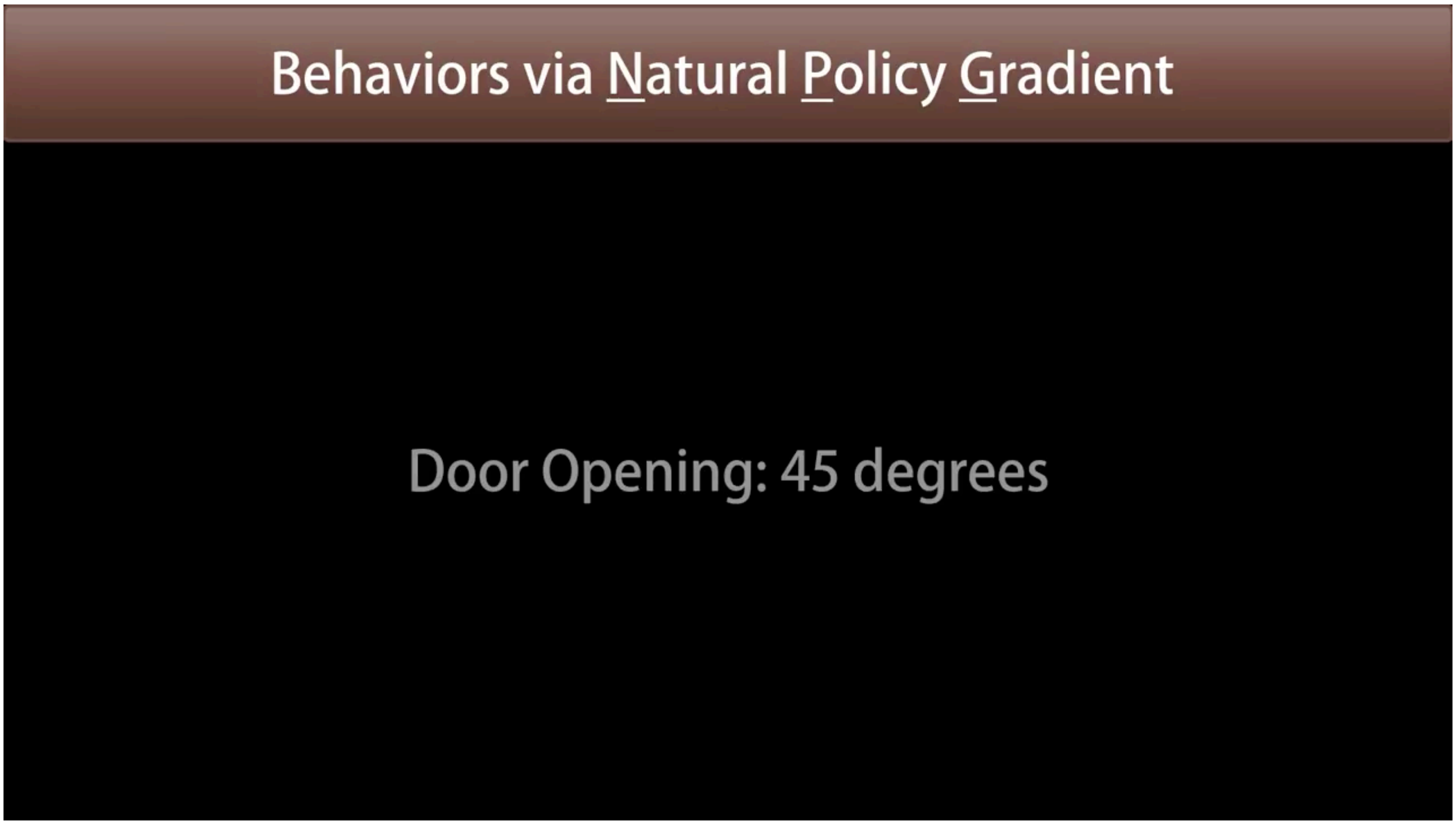
Problems and Solutions in AI Safety



Gokul Swamy

Goya, 1799

The Good

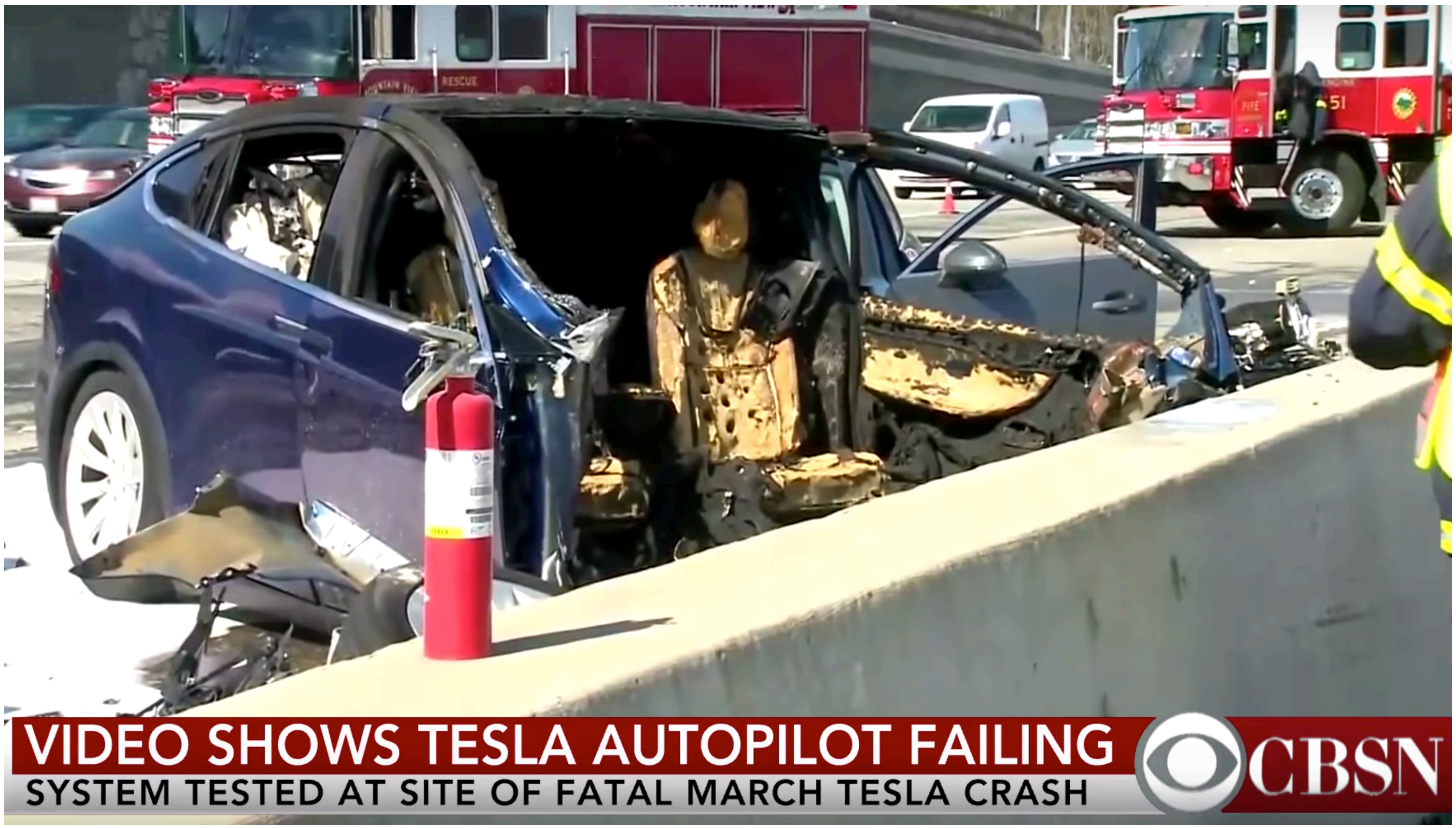


Zhu, Gupta et al.

The Bad

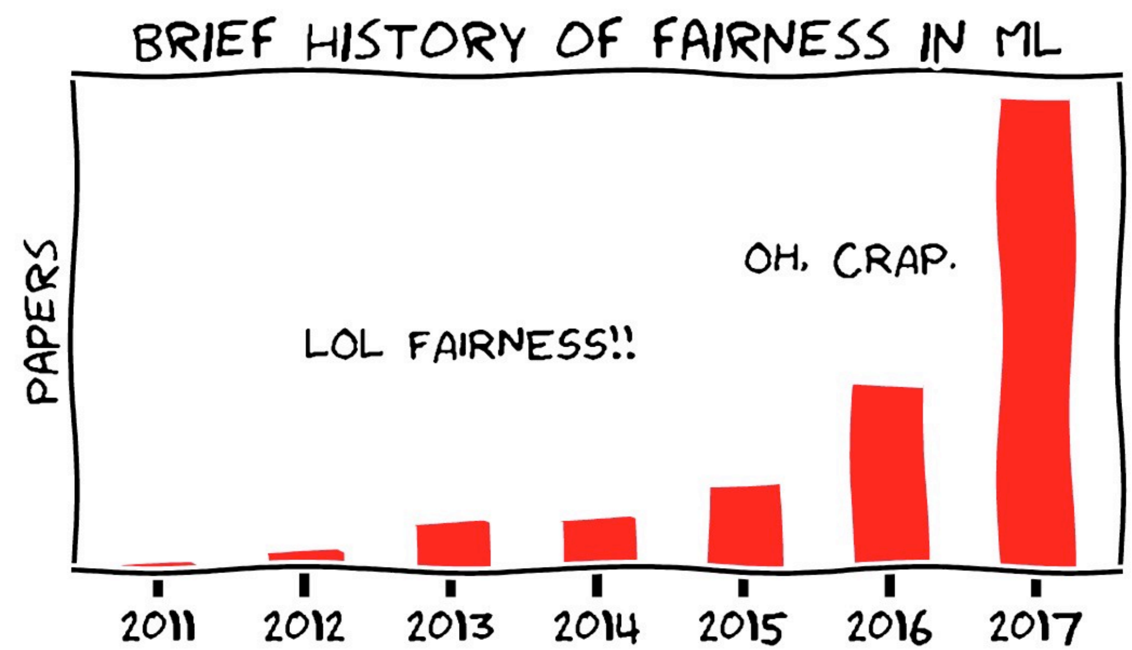
Amodei et al.

The Ugly

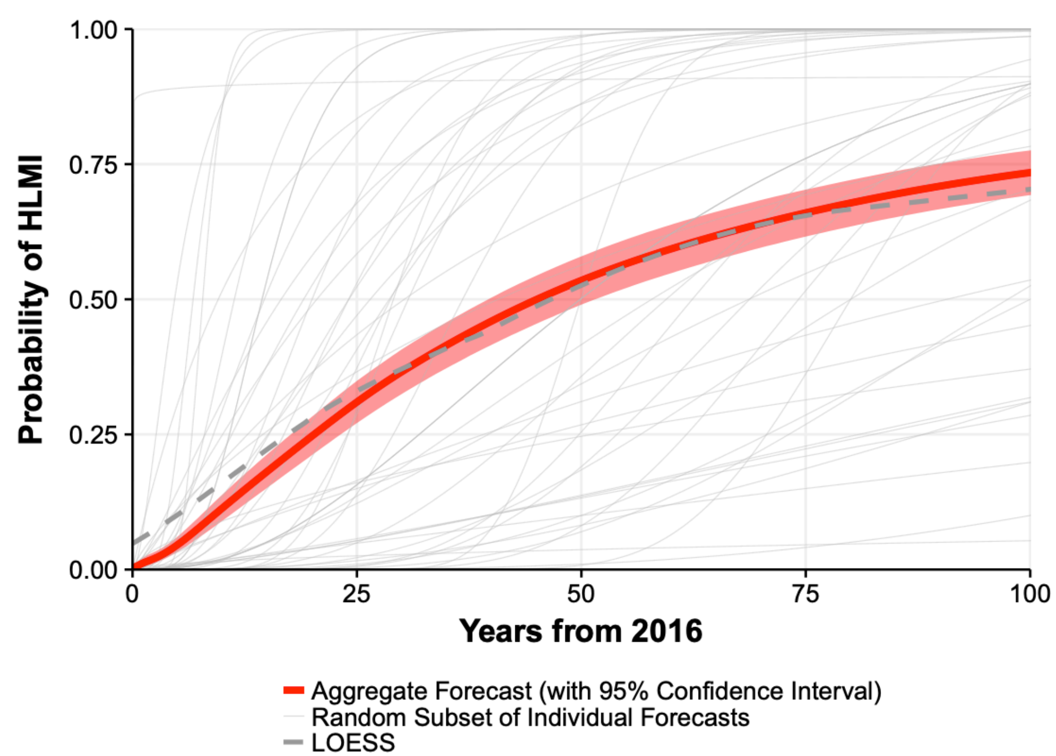

VIDEO SHOWS TESLA AUTOPILOT FAILING
SYSTEM TESTED AT SITE OF FATAL MARCH TESLA CRASH
CBSN

Why talk about this now?

- Increased Use:



- Grace et al.:



Why talk about this now?

"If you say, 'Fetch the coffee', it can't fetch the coffee if it's dead. So if you give it any goal whatsoever, it has a reason to preserve its own existence to achieve that goal."

\- **Stuart Russell**

"If a superior alien civilization sent us a text message saying, 'We'll arrive in a few decades,' would we just reply, 'OK, call us when you get here — we'll leave the lights on'?"

\- **Stuart Russell**

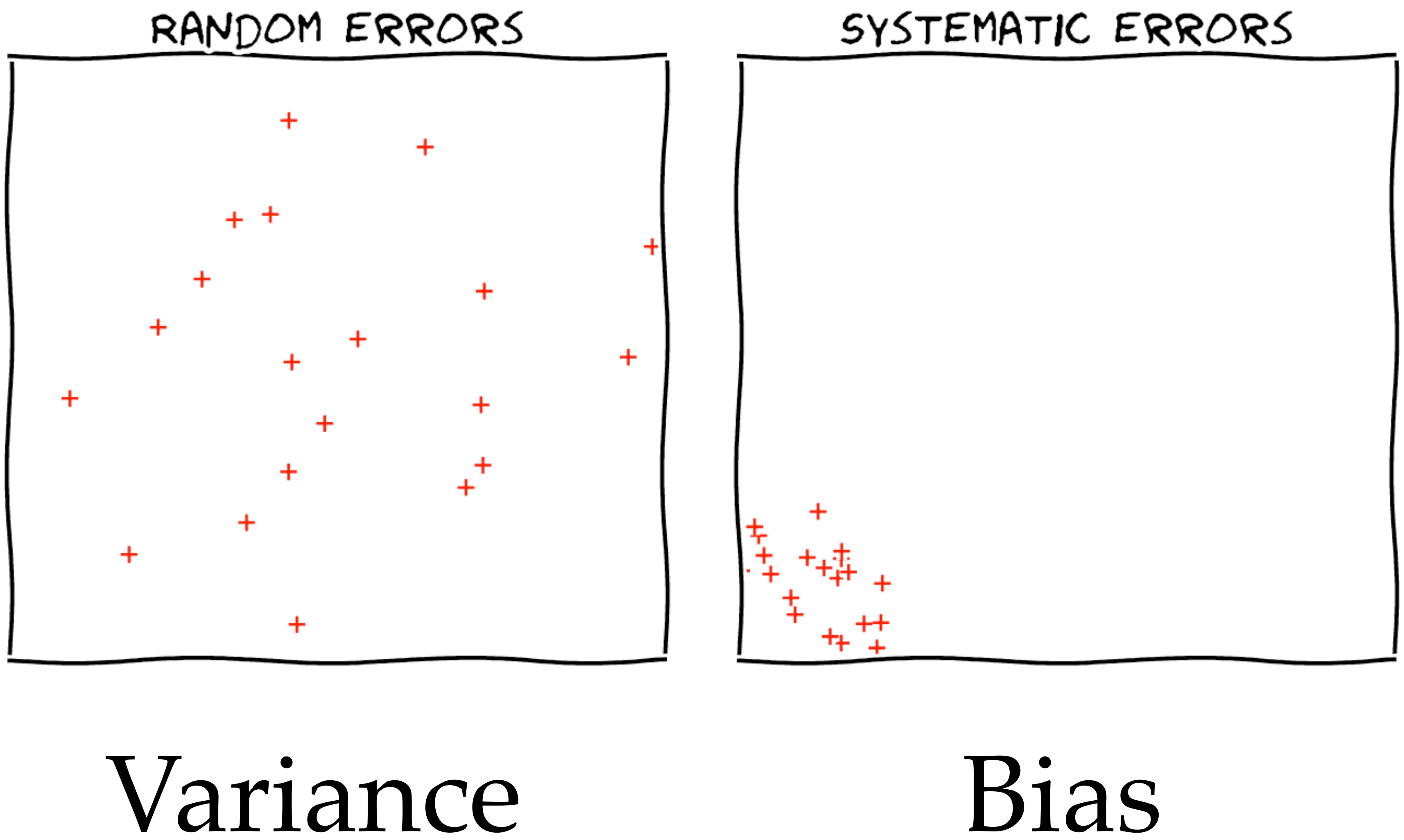
3 Problems in AI Safety

- Arranged in order of decreasing urgency:
  - Problem 1: Algorithmic Bias
  - Problem 2: Safe Exploration
  - Problem 3: Value Alignment

Disclaimers

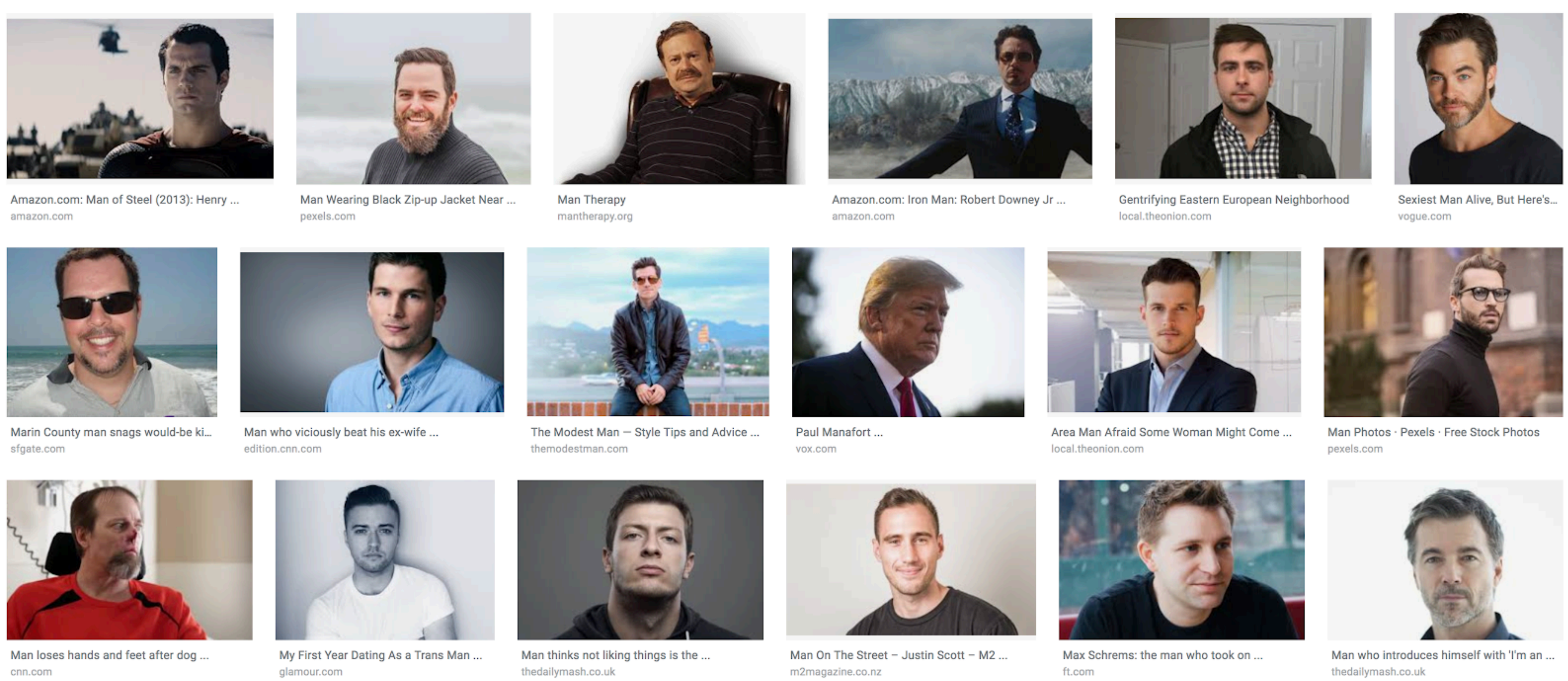
- I'm going to focus on the key ideas behind the papers we're going to talk about today rather than the mathematical details
  - Please read them yourself if interested in precise justifications
- Most of the research here was done by people at Cal
  - Don't overfit to a single set of viewpoints
- I have not timed this lecture
  - So let's begin!

P1: What is bias?



Where does bias come from?

- From datasets:
  - Google search for “man”



Where does bias come from?

- From social dynamics:

Why is algorithmic bias particularly bad?

- Because a result is produced by a computer, people believe it more

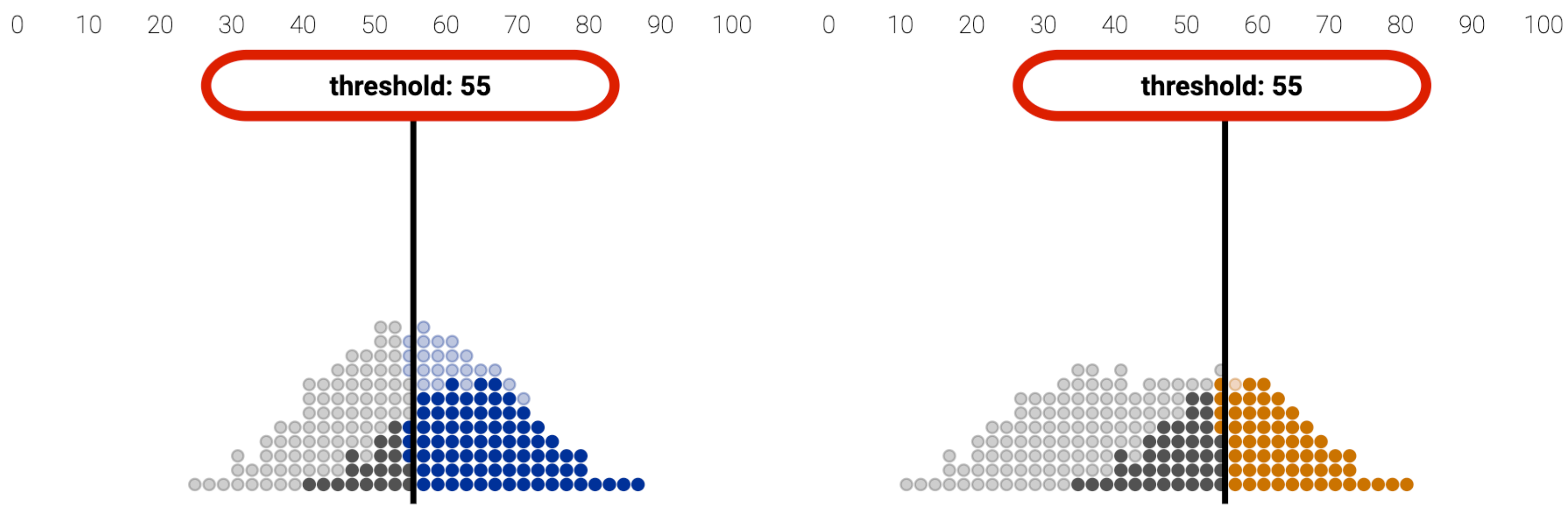
- Amazon hiring tool:

Problem Setup

- **Protected Class**: gender, race, sexual orientation, …
- Using Hardt et al.'s terminology.
- Running example: Granting life insurance policy based on gender and age.

What sort of fairness criterion do we want?

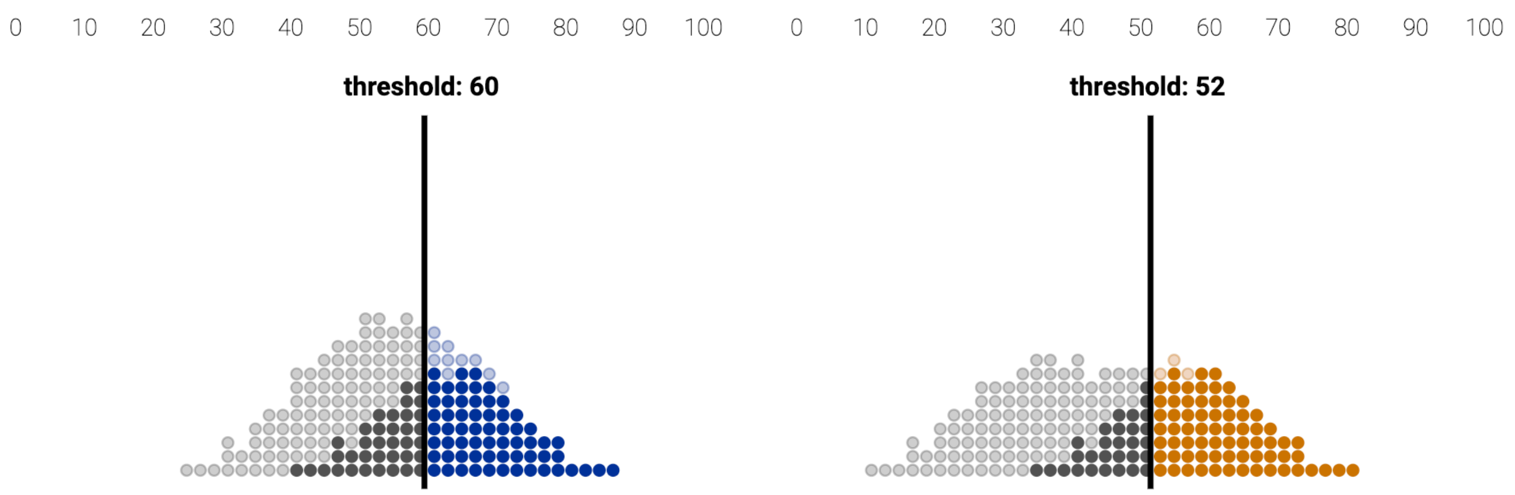
- **Group Unaware**: Same threshold across groups
  - Problem: Women on average live longer than men



  - More men and fewer women have been granted loans than they should

What sort of fairness criterion do we want?

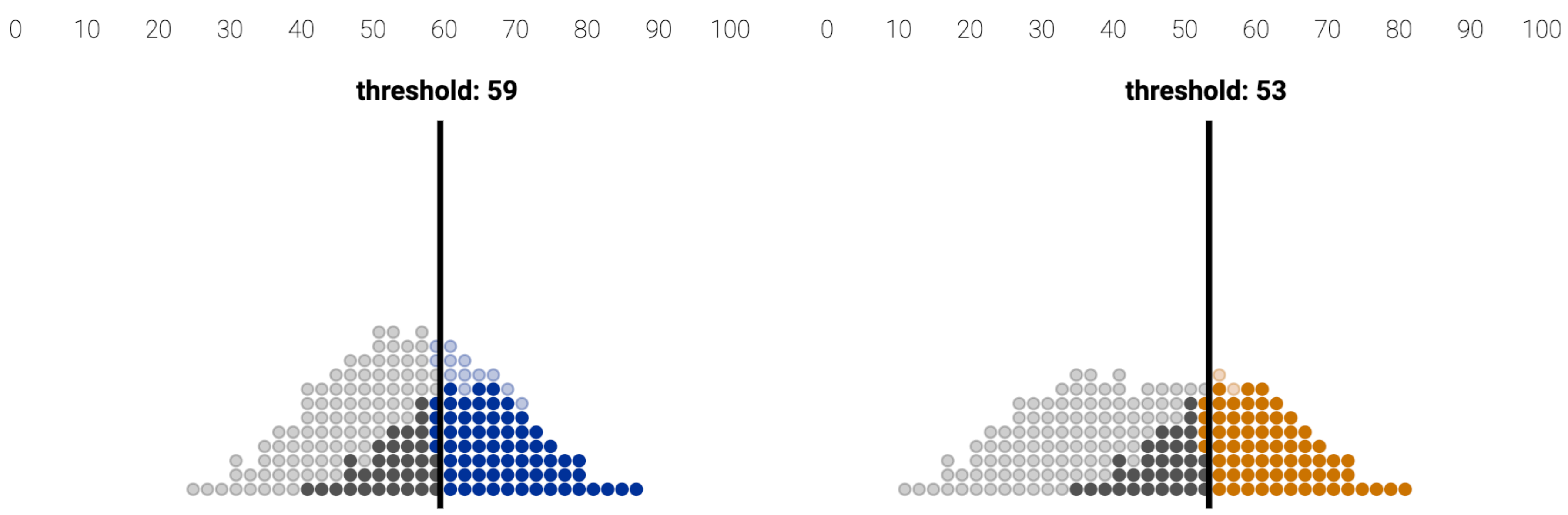
- **Demographic Parity**: Same positive rate across groups
  - Same proportion of colored dots



  - However, this leads to more men who would make payments getting denied than women in the same situation.
    - Ignores difference in default rates across groups

What sort of fairness criterion do we want?

- **Equal Opportunity**: Same *true* positive rate across groups
  - Same proportion of dark colored dots



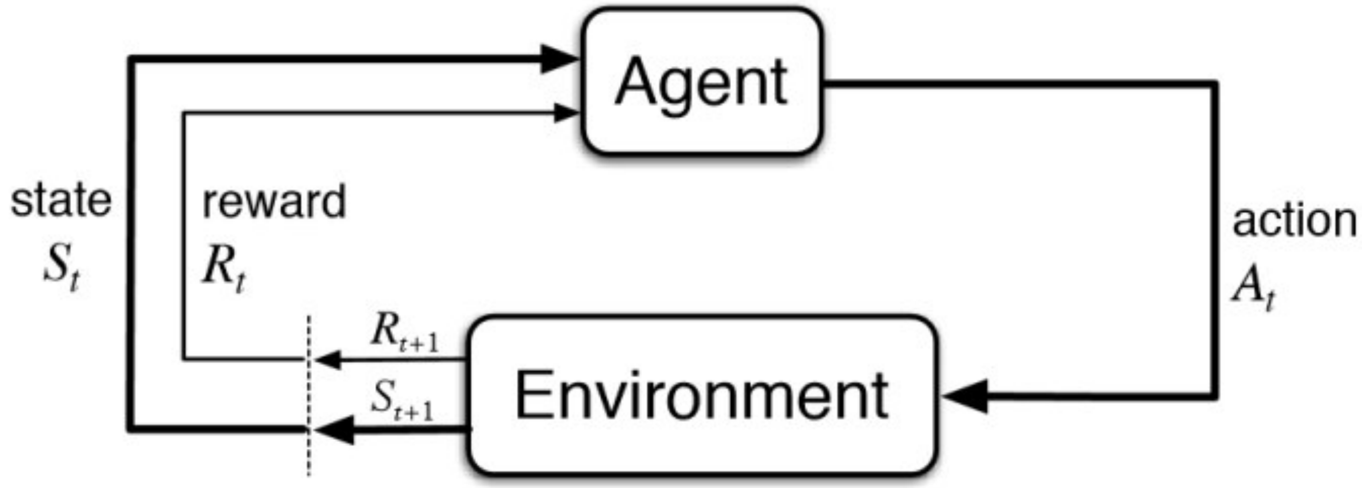
  - Conditioned on knowing someone's chance of making payments, their gender provides no more information.

P2: Safe Exploration

- When we have robots in the real world, we want them to be safe
  - We want them to not mess up environments
  - We want them to interact with people in ways that make them feel comfortable

RL Framework

- RL can be considered a generalization of supervised learning:

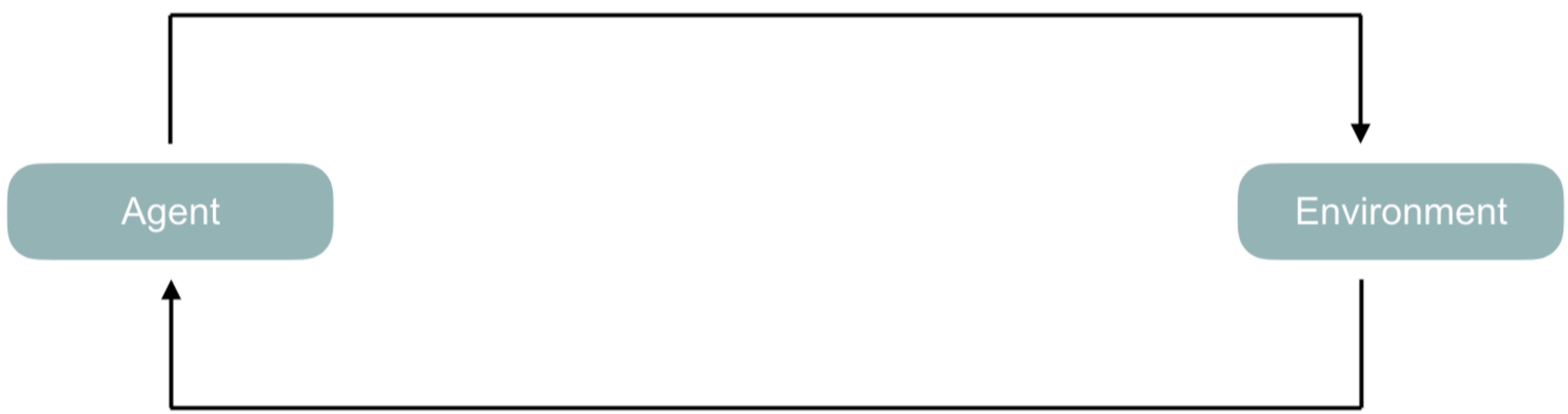


RL Definitions

- **Environment**: The world in which our problem is set up. The environment updates according to **dynamics**
- **State**: All the aspects of the environment at a particular time that are relevant to the problem we're trying to solve
- **Agent**: Can take actions to influence the state of the world
- **Policy**: How our agent decides to act given the state of the world. A distribution over actions given state.
- **Trajectory**: List of state-action tuples generated by our interaction with env.

RL Formalized

Takes action a_t according to $\pi(a_t, s_t)$



Give us s_{t+1} by sampling from $T(s_{t+1}|s_t, a_t)$ and $r = R(s_t, a_t)$

Value and Q Functions

- Discounted sum of future rewards:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- The average of this defines the "value" of a state:

$$V^\pi(s) = \mathbb{E}_\pi \left[R_t | s_t = s \right]$$

- We can break this down even further to actions:

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[R_t | s_t = s, a_t = a \right]$$

Inverse Reinforcement Learning

- What if instead of learning to act to maximize reward, we want to learn a reward function that would, when maximized, lead to demonstrated behavior?
  - This is **inverse reinforcement learning**
- Traditional recipe (Abbeel and Ng):
  - 1) Determine some higher-level features of state that someone would likely care about: $\theta(s)$
  - 2) Write your reward function as $R(s) = w^T \theta(s)$
  - 3) Fit weights such that actions taken maximize reward
- This guarantees behavior that matches feature counts of demonstrations in expectation

Maximum Entropy IRL

- The previous recipe requires the demonstrator to be exactly optimal
  - People are very rarely perfect
- Instead, we can assume people are Boltzmann Rational or "**noisily rational**":

$$\mathbb{P}((s_i, a_i)|R) = \frac{1}{Z_i}\exp\{\alpha Q^*(s_i, a_i, R)\}$$

Why is this assumption ok?

- Most conservative assumption we can make - we only assume what we have to make sure feature counts match
    - *The exponential distribution maximizes entropy given a constraint on the first moment*
- People definitely don't act like this though
    - "You don't open the trunk of your car to get into the driver's seat with some small probability, you just don't" - Stuart Russell
    - "All models are wrong but some are useful" - George Box

Recovering Reward Functions

$$\mathbb{P}(\tau|R) = \prod_{i=1}^{n} \mathbb{P}((s_i, a_i)|R)$$

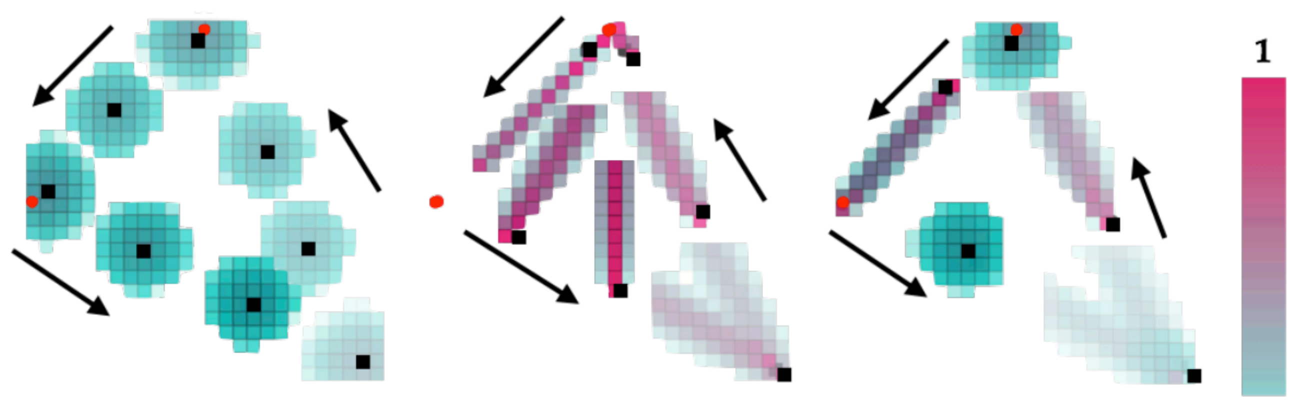
$$\mathbb{P}(\tau|R) = \frac{1}{Z}\exp\{\alpha \sum_{i=1}^{n} Q^*(s_i, a_i, R)\}$$

$$\mathbb{P}(R|\tau) = \frac{\mathbb{P}(\tau|R)\mathbb{P}(R)}{\mathbb{P}(\tau)} = \frac{1}{Z}\exp\{\alpha \sum_{i=1}^{n} Q^*(s_i, a_i, R)\}\mathbb{P}(R)$$

***Key Point**: We get a distribution over reward functions*

Probabilistically Safe Robot Planning with Confidence-Based Human Predictions

- Problem: We want to make sure robots don't hit people when they are moving
- Fisac et al.'s Key Idea: Online estimate beta (same as alpha from before) so we can be more conservative when necessary



(a) β low confidence (b) β high confidence (c) β inference

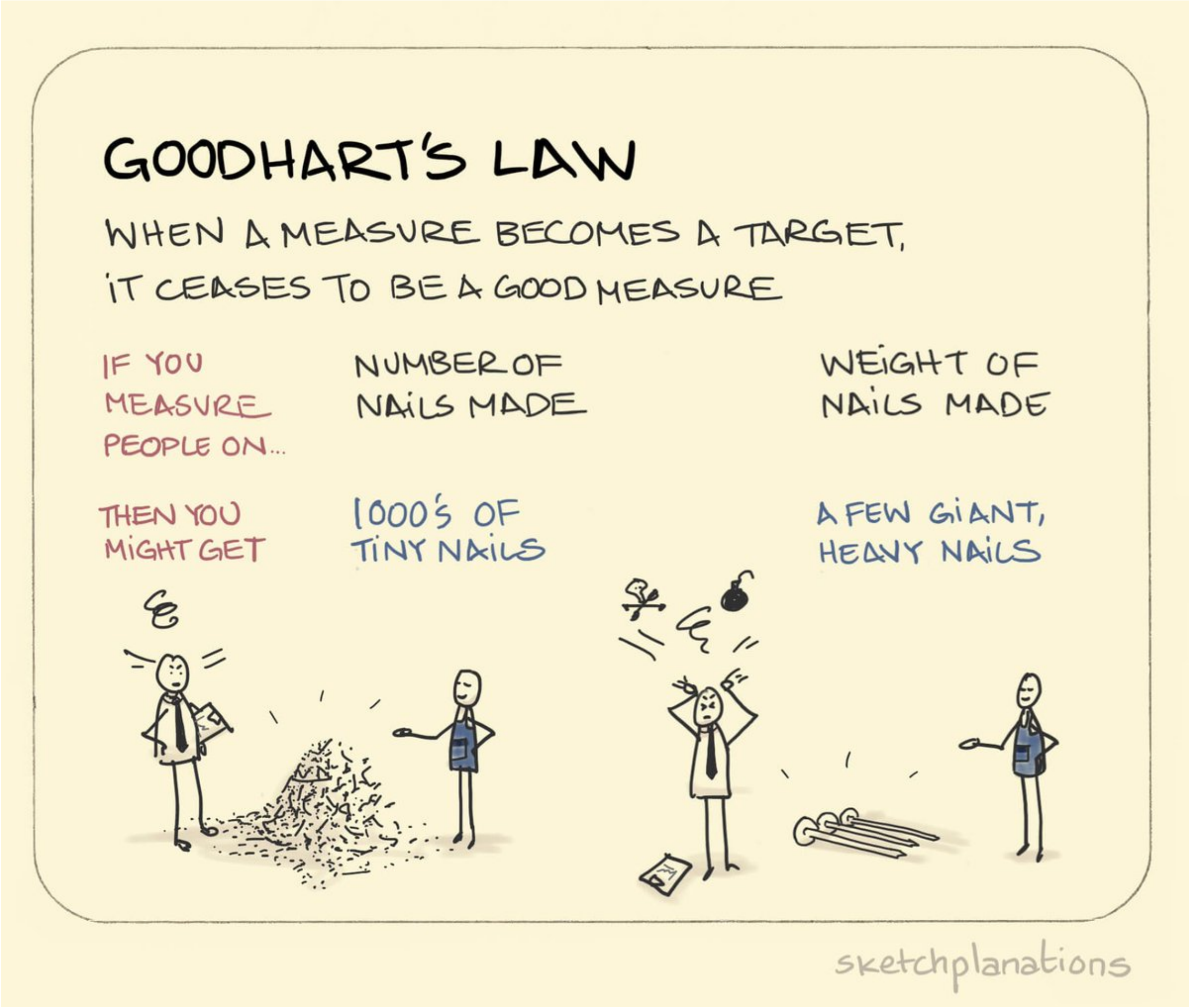
- Anca's Explanation: https://youtu.be/_VceNn8ZWAg?t=18105

Value Alignment

- Value Alignment: AI agents doing what we *actually want*
- Counter-examples:
  - Folklore: King Midas
  - Media: Sorcerer's Apprentice: https://www.youtube.com/watch?v=3REmfMKhlO0
  - Robotics: Asking a cleaning robot to pick up as much dust as possible.

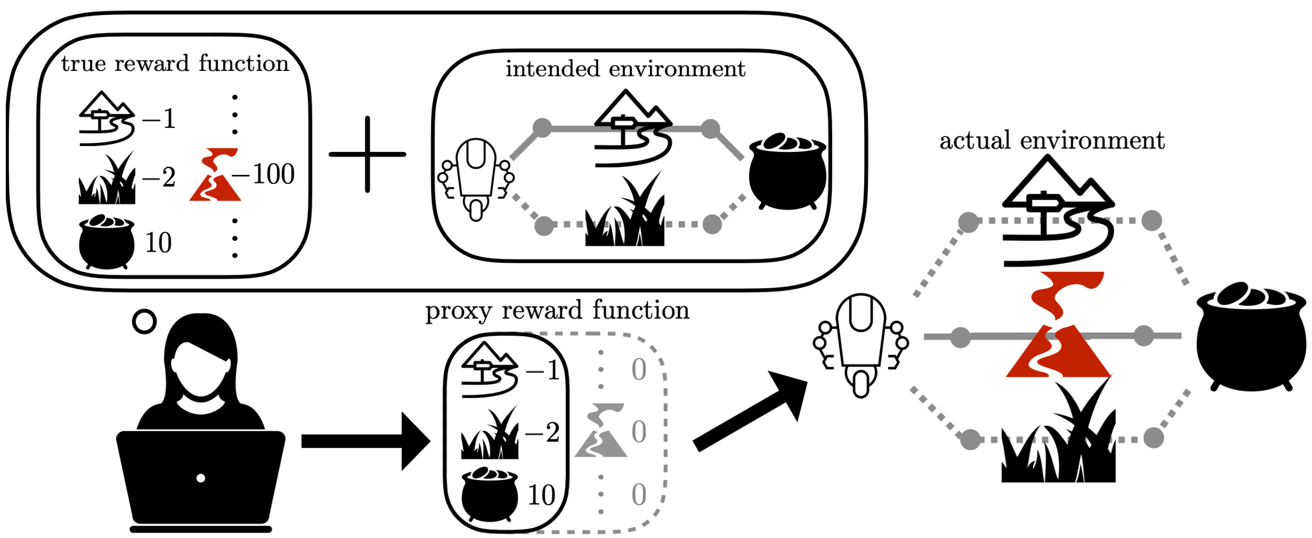
What was the problem here?

- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."

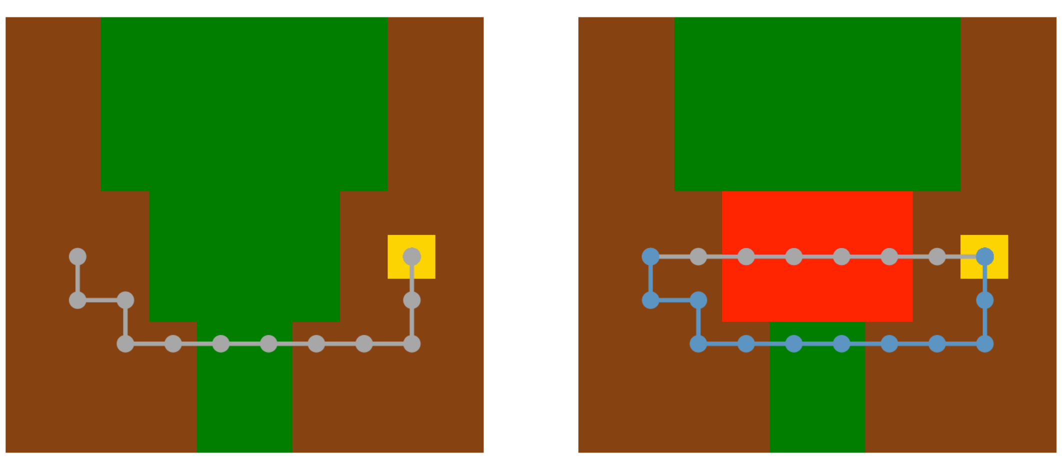


Inverse Reward Design

- **Key Idea**: Treat given reward function as observation about true reward function in designer's head

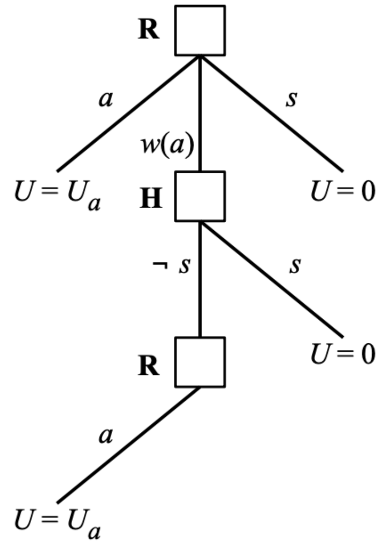


$$P(w = w^* | \tilde{w}, \tilde{M}) \propto \frac{\exp\left(\beta w^\top \tilde{\phi}\right)}{\tilde{Z}(w)} P(w)$$



The Off-Switch Game

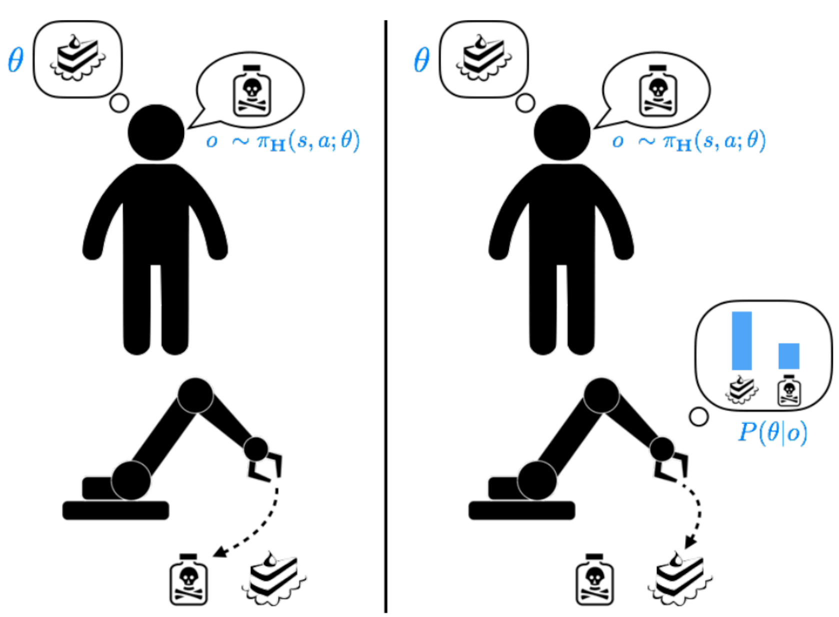
- Some human error will always slip through
  - We want our systems to be **corrigible** - we can stop them if needed
- **Key Idea**: For systems to be corrigible, they need to have some uncertainty about their utility functions

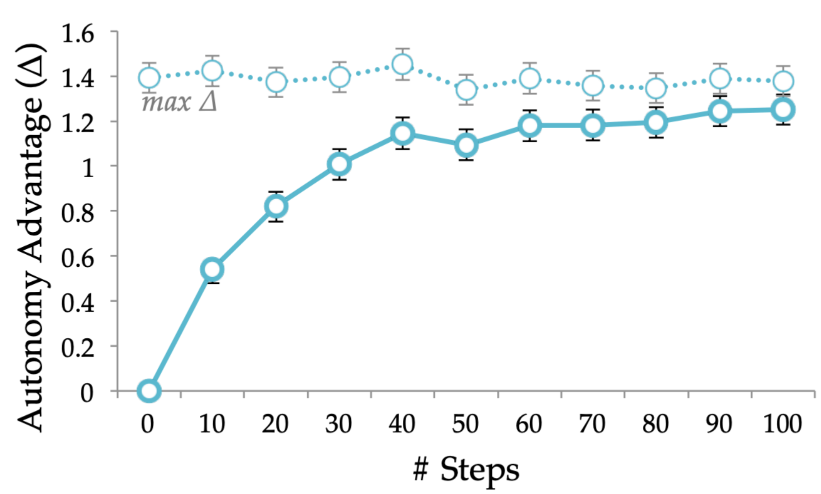


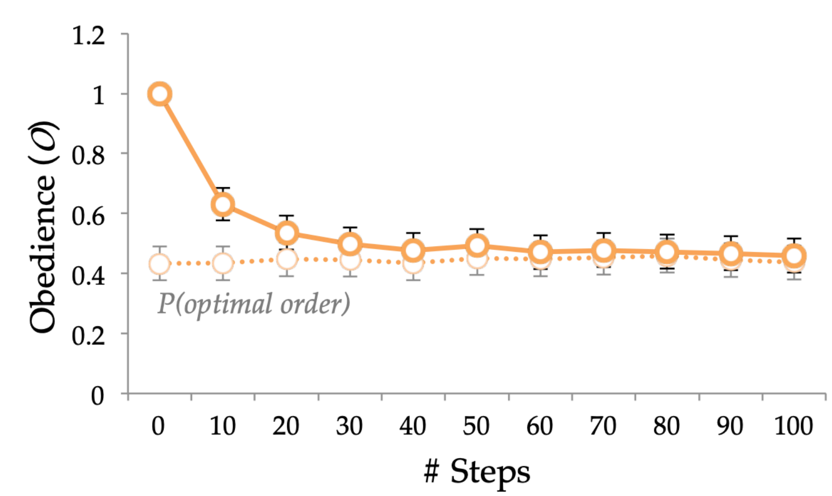
$$\pi^{\mathbf{H}}(U_a) = \begin{cases} 1 & U_a \geq 0 \\ 0 & o.w. \end{cases}$$

Should robots be obedient?

- We don't always want robots to listen to people
  - Consider a self-driving car dropping off a truant to school
- **Key Idea**: A robot should intelligently decide whether to listen to a person
  - See paper for details of how this applies to noisily rational people

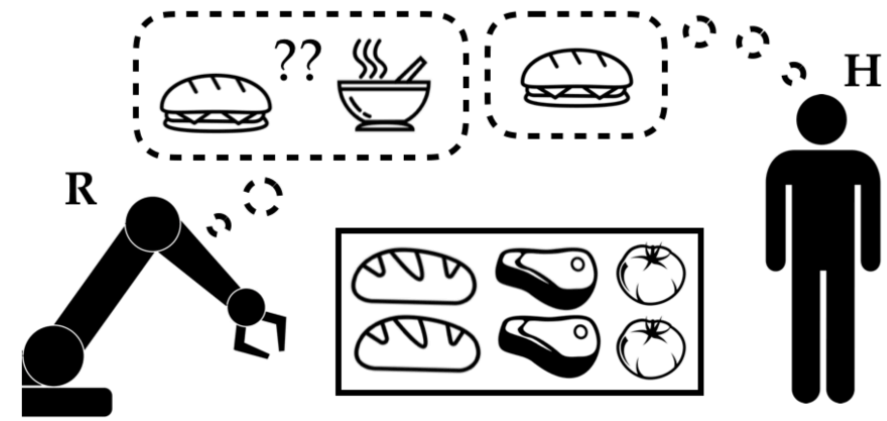


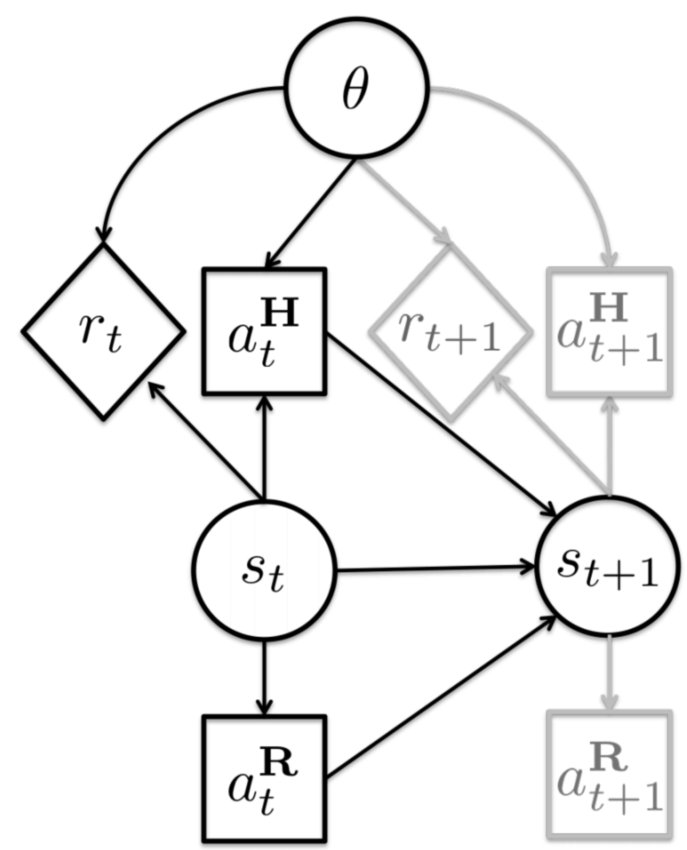




Cooperative Inverse Reinforcement Learning

- Consider a cooperative game with 2 players: a human and a robot
  - Cooperative so they receive they same reward
  - However, only human knows reward parameters
  - Robot is trying to use IRL to recover them from human behavior
- This formulation incentives active teaching and active learning





Takeaways

- As AI becomes increasingly integrated into our world, we need to take a closer look at the implications of the technologies we're using
- In the short term, we need to make sure our algorithms are not as biased as the data they are fed
- In the middle term, we need to make sure robots are cognizant of the people we are interacting with
- In the long term, we need to make sure our AI agents use uncertainty to be human-compatible